УДК 519.237.8+510.22

Д.А. Вятченин
Объединенный институт проблем информатики НАН Беларуси, г. Минск
viattchenin@mail.ru

Применение нечетких чисел для обоснования кластеров в методах нечеткой кластеризации*

В статье рассматривается метод определения оптимального числа кластеров в нечетком c-разбиении, основанный на построении интервала значений наиболее возможного числа классов с использованием нечетких чисел. Предложена модификация FCM-CV-алгоритма и приводится результат вычислительного эксперимента.

Введение

Теория нечетких множеств, предложенная в 1965 году Л.А. Заде [1], нашла свое применение практически в любой области научных исследований, в том числе и при разработке новых методов автоматической классификации, именуемой также кластерным анализом, численной таксономией или распознаванием образов с самообучением [2], в связи с чем Л.А. Заде отмечал, что «глубинная связь между теорией нечетких множеств и распознаванием образов основана на том обстоятельстве, что большинство реальных классов размыты по своей природе в том смысле, что переход от принадлежности к непринадлежности для этих классов скорее постепенен, чем скачкообразен» [3]. В работах [4-6] подробно рассматриваются предложенные различными исследователями методы нечеткой кластеризации эвристического, оптимизационного и иерархического направлений. Оптимизационные методы нечеткого подхода к решению задач кластеризации являются наиболее распространенными, однако при обращении к указанным кластер-процедурам возникает проблема обоснования числа классов, для решения которой традиционно используются различные показатели оптимальности числа нечетких кластеров. Вместе с тем, при большом количестве предположений о числе классов использование этих показателей также затруднительно, и задача состоит в построении множества наиболее возможного числа нечетких кластеров.

Целью предпринятого исследования является решение поставленной задачи, в основе которого лежит использование аппарата нечетких чисел. В работе предлагается метод построения допустимого множества значений наиболее возможного числа нечетких кластеров в искомом нечетком c-разбиении, основанный на построении нечетких величин, исходя из экспертных оценок или результатов разведочного анализа данных, а также рассматривается соответствующая модификация предложенного в [7] FCM-CV-алгоритма. Приводятся результаты вычислительного эксперимента и формулируются предварительные выводы об эффективности предложенного подхода.

---

*Исследования проводились при поддержке гранта Президиума Национальной академии наук Беларуси в соответствии с Постановлением № 157 Бюро Президиума НАН Беларуси от 11.04.2007.

Оптимизационное направление решения нечеткой модификации задачи кластеризации

В нечетких методах автоматической классификации, объединяемых в оптимизационное направление, нечеткая кластеризация понимается как разбиение классифицируемой совокупности объектов $X = \{x_1,...,x_n\}$ на семейство его нечетких множеств, так что в качестве входного параметра в существующих нечетких методах оптимизационного направления автоматической классификации задается число нечетких кластеров c, причем при данном подходе под нечетким кластером может пониматься любое нечеткое множество, определенное на универсуме. Нечеткие множества $A^l, l = 1,\dots,c$ с соответствующими функциями принадлежности $\mu_{1i},\dots,\mu_{ci}$ каждого объекта x_i, определенные на универсуме $X = \{x_1,...,x_n\}$, образуют нечеткое c-разбиение, иногда называемое также нечетким разбиением в смысле Э.Г. Распини [3], если для каждого объекта $x_i \in X$ выполняется условие $\sum_{l=1}^{c}\mu_{li} = 1$, и нечеткая модификация задачи автоматической классификации в экстремальной постановке заключается в нахождении экстремума некоторого функционала $Q(P)$ на множестве Π всех возможных нечетких c-разбиений P классифицируемого множества объектов X, что описывается формулой $Q(P) \to \underset{P\in\Pi}{extr}$.

Если исходные данные представлены матрицей вида «объект-объект» $d_{n\times n} = [d(x_i, x_j)], i = 1,\dots,n, j = 1,\dots,n$, элементами которой являются попарные коэффициенты различия между объектами, то для классификации объектов исследуемой совокупности используются так называемые реляционные процедуры, основанные на минимизации соответствующих функционалов. Примером такого функционала может послужить критерий М. Рубенса [8]

$$Q^{I}_{Ro}(P) = \sum_{l=1}^{c}\sum_{i=1}^{n}\sum_{j=1}^{n}\mu_{li}^2\mu_{lj}^2 d(x_i, x_j\} , \qquad (1)$$

процедура минимизации которого именуется в литературе MND2-алгоритмом [8] или FNM-алгоритмом [6]. В выражении (1) символом $d(x_i, x_j)$ обозначается коэффициент различия между объектами x_i и x_j, $i, j = 1,\dots,n$ исследуемой совокупности X, $card(X) = n$, а μ_{li} – значение принадлежности i-го объекта l-му нечеткому кластеру.

В случае же представления исходных данных в виде матрицы «объект – свойство» $\hat{X}_{n\times m} = [\hat{x}_i^t], i = 1,\dots,n, t = 1,\dots,m$, задача классификации заключается в минимизации критерия, в который введена некоторая метрика, и типичным примером подобных функционалов является критерий, предложенный Дж. Данном [9] и позже обобщенный Дж. Беждеком [10]

$$Q^{II}_{DB}(P, \overline{\mathrm{T}}) = \sum_{l=1}^{c}\sum_{i=1}^{n}\mu_{li}^{\gamma}\left\|x_i - \overline{\tau}^l\right\|^2 , \qquad (2)$$

процедура минимизации которого широко известна под обозначением FCM-алгоритма. Критерий (2), где символом $\overline{\tau}^l$ обозначен прототип l-го нечеткого кластера, а символом γ – задаваемый исследователем коэффициент нечеткости классификации, такой, что $1 < \gamma < \infty$, послужил основой для целого семейства функционалов и соответствующих им нечетких кластер-процедур, которые подробно рассматриваются в [4].

В силу того, что c является параметром любой оптимизационной кластер-процедуры, одной из главных проблем при использовании оптимизационных методов является определение «реального» числа c нечетких кластеров, на которые «расслаивается» исследуемая совокупность, или, иными словами, проблема обоснования числа кластеров, встающая наиболее остро, когда исследователю число классов c вообще неизвестно. Для решения этой проблемы были предложены различные показатели, характеризующие получаемое при использовании того или иного алгоритма нечеткое c-разбиение $P^* = \{A^1, \ldots, A^c\}$. В частности, для FCM-алгоритма и его модификаций различными исследователями был введен ряд показателей, наиболее известными из которых являются

– коэффициент разбиения

$$V_{pc}(P) = \frac{1}{n}\sum_{l=1}^{c}\sum_{i=1}^{n}\mu_{li}^2, \qquad (3)$$

предложенный Дж. Данном [11] и для которого решение задачи определения оптимального числа классов в P^* отыскивается в виде $\max\limits_{c}\left(V_{pc}(P)\right), c = 2, \ldots, n-1$;

– энтропия разбиения

$$V_{pe}(P) = -\frac{1}{n}\sum_{l=1}^{c}\sum_{i=1}^{n}\left|\mu_{li}\cdot \ln \mu_{li}\right|, \qquad (4)$$

предложенная Дж. Беждеком, М.П. Уиндхемом и Р. Эрлихом [12], так что оптимальному числу классов в P^* соответствует $\min\limits_{c}\left(V_{pe}(P)\right), c = 2, \ldots, n-1$;

– индекс разделимости

$$V_{si}(P) = \frac{\frac{1}{n}\sum_{l=1}^{c}\sum_{i=1}^{n}\mu_{li}^2\left\|x_i - \tau^l\right\|^2}{\min\limits_{l \neq k}\left\|\tau^l - \tau^k\right\|^2}, \qquad (5)$$

предложенный Х.Л. Хи и Ж. Бени [13], где оптимальное число классов в P^* определяется, исходя из условия $\min\limits_{c}\left(V_{si}(P)\right), c = 2, \ldots, n-1$. Для других нечетких кластер-процедур оптимизационного направления также предлагается ряд показателей оптимальности числа классов в нечетком c-разбиении – в частности, в [14] рассматриваются показатели оптимальности нечетких кластеров при разбиении исследуемой совокупности с помощью FNM-алгоритма. Для всех показателей числа классов в нечетком c-разбиении решение задачи определения оптимального числа классов в искомом нечетком разбиении определяется общим выражением

$$\underset{c}{extr}\left(V_c(P)\right), \quad c = 2, \ldots, n-1, \qquad (6)$$

где символом $V_c(P)$ обозначен какой-либо показатель. Таким образом, необходимым является проведение серии экспериментов при различных значениях числа классов c, для чего оказывается необходимым построение множества $C = \{c_*, \ldots, c^*\}$ наиболее возможных значений числа классов в искомом нечетком c-разбиении P^*, где c_* – наименьшее, а c^* – наибольшее из значений множества C.

В [15] было предложено объединить FCM-алгоритм с процедурой вычисления соответствующего показателя оптимальности числа нечетких классов $V_c(P)$ в искомом нечетком c-разбиении P^* в рамках одной процедуры, параметрами которой являются значения наименее возможного c_* и наиболее возможного c^* числа нечетких кла-

стеров в искомом нечетком c-разбиении P^*. Вместе с тем, предложенная в [15] процедура предъявляет достаточно высокие требования к оперативной памяти ПЭВМ, особенно при обработке совокупностей сравнительно большого объема, и в [7] была предложена модификация предложенной в [15] процедуры, получившая, от англоязычных терминов fuzzy c-means и cluster validity, название FCM-CV-алгоритма. Рассмотренный в предпринятом исследовании метод построения множества наиболее возможного числа нечетких кластеров в искомом нечетком c-разбиении P^* позволяет модифицировать FCM-CV-алгоритм, и соответствующая модификация процедуры будет рассмотрена в процессе дальнейшего изложения.

Построение множества значений возможного числа нечетких кластеров на основе нечетких чисел

Для дальнейшего рассмотрения представляется необходимым напомнить определения понятий нечеткой величины, нечеткого интервала и нечеткого числа. Если некоторое нечеткое множество V определено на множестве действительных чисел, то есть представляет собой отображение $\Re \to [0,1]$, то V именуется нечеткой величиной. Каждое значение $x_i \in \Re$ будет называться модальным значением нормальной нечеткой величины V, если x_i является элементом ядра $Core(V)$ нечеткой величины V, то есть $\mu_V(x_i) = 1$. В случае, если $card(Core(V)) = 1$, то V является унимодальной нечеткой величиной, в случае же, когда $card(Core(V)) > 1$, нечеткая величина V называется мультимодальной [16].

Пусть L или R – невозрастающая функция $\Re^+ \to [0,1]$, такая, что $L(0) = R(0) = 1$ и $\forall x_i > 0, L(x_i) < 1, \forall x_i < 1, L(x_i) > 0$; $L(1) = 0$ либо имеет место $L(x_i) > 0, \forall x_i$ и $L(+\infty) = 0$. Тогда нечеткое множество V называется нечетким интервалом LR-типа с функцией принадлежности $\mu_V(x_i)$, определяемой формулой

$$\mu_V(x_i) = \begin{cases} L\left(\frac{\underline{m} - x_i}{a}\right), & x_i \le \underline{m} \\ 1, & \underline{m} \le x_i \le \overline{m} \\ R\left(\frac{x_i - \overline{m}}{b}\right), & x_i \ge \overline{m} \end{cases}, \quad (7)$$

где $\underline{m}$ называется нижним модальным значением V, $\overline{m}$ – верхним модальным значением, а параметры $a > 0$ и $b > 0$ называются левым и правым коэффициентами нечеткости соответственно. Таким образом, нечеткий интервал LR-типа может быть представлен в виде четверки параметров, что записывается в виде $V = (\underline{m}, \overline{m}, a, b)_{LR}$.

Если $V = (\underline{m}, \overline{m}, a, b)_{LR}$ является нечетким интервалом LR-типа, то при условии совпадения нижнего и верхнего модальных значений $\underline{m} = \overline{m} = m$ нечеткий интервал LR-типа V именуется нечетким числом LR-типа с функцией принадлежности, определяемой в соответствии с выражением

$$\mu_V(x_i) = \begin{cases} L\left(\frac{m - x_i}{a}\right), & при \quad x_i \le m \\ R\left(\frac{x_i - m}{b}\right), & при \quad x_i \ge m \end{cases}, \quad (8)$$

где m называется модальным значением нечеткого числа LR-типа, символически записываемого в виде $V = (m, a, b)_{LR}$, а a и b – левым и правым индексами нечеткости соответственно.

При использовании оптимизационных методов нечеткой кластеризации число классов c в искомом нечетком c-разбиении P^* может определяться с помощью результатов разведочного анализа данных, к примеру, при анализе диаграммы рассеивания [2] или экспертных оценок. Пусть $\mathcal{C}_e = \{(\text{€}_e, \mu_{\mathcal{C}_e}(\text{€}_e)) \,|\, 2 \le \text{€}_e \le n, \forall e = \overline{1,k}\}$ – множество одноточечных нечетких множеств, выражающих мнения k экспертов о числе c кластеров в искомом нечетком c-разбиении P^*, так что предполагаемому e-м экспертом числу €_e нечетких кластеров в P^* соответствует степень уверенности $\mu_{\mathcal{C}_e}(\text{€}_e) \in (0, 1]$. Для каждого $e = 1, \ldots, k$ значение €_e может рассматриваться как модальное значение нечеткого числа, и для каждого из строящихся нечетких чисел $V_e = (m_e, a_e, b_e)_{LR}$, $e = 1, \ldots, k$ с модальными значениями $m_e = \text{€}_e$ соответственно значения $\mu_{V_e}(l)$, $e = 1, \ldots, k$ будут полагаться равными нулю в точках $c = 1$ и $c = n$, так что $\mu_{V_e}(1) = \mu_{V_e}(n) = 0$, $\forall e \in \{1, \ldots, k\}$, левый коэффициент нечеткости будет определяться выражением $a_e = (\text{€}_e - 1)$, а правый – $b_e = (n - \text{€}_e)$, так что носителем $Supp(V_e)$ нечеткого числа V_e, $e = 1, \ldots, k$ будет открытый интервал $(\text{€}_e - a_e, \text{€}_e + b_e) = (1, n)$, а $\mu_{V_e}(l)$ будет определяться выражением

$$\mu_{V_e}(l) = \begin{cases} L_e\left(\dfrac{\text{€}_e - l}{a_e}\right), & при \quad l \le \text{€}_e \\ R_e\left(\dfrac{l - \text{€}_e}{b_e}\right), & при \quad l \ge \text{€}_e \end{cases}, \tag{9}$$

где индекс e функций L и R подчеркивает, что они могут выбираться для каждого нечеткого числа $V_e = (m_e, a_e, b_e)_{LR}$, $e = 1, \ldots, k$ отдельно. Таким образом, если на координатной плоскости по оси абсцисс откладывать число классов $l \in (1, n)$, а по оси ординат – значения функции принадлежности $\mu_{V_e}(l)$ нечеткого числа V_e, то функция представления формы нечеткого числа V_e, $e = 1, \ldots, k$ будет иметь вид кривой, достигающей максимума в точке с координатами $l = \text{€}_e$, $\mu_{V_e}(l) = 1$. При построении нечетких чисел V_e, $e = 1, \ldots, k$ для того, чтобы каждое V_e, $e = 1, \ldots, k$ описывалось бы непрерывной функцией принадлежности $\mu_{V_e}(l)$, подразумевается, что $l \in (1, n) \subset \Re$.

Далее следует построить нечеткую величину V с непрерывной функцией принадлежности $\mu_V(l)$, для чего ко всем $V_e = (m_e, a_e, b_e)_{LR}$, $e = 1, \ldots, k$ можно применить операцию объединения, причем в данном случае операция объединения понимается в широком смысле, то есть может быть осуществлена с помощью какой-либо выбранной исследователем S-нормы [17]. После построения V выбирается порог $\alpha = \min\limits_e \mu_{\mathcal{C}_e}(\text{€}_e)$, для которого строится $V_{(\alpha)} = \{(l, \mu_{V_{(\alpha)}}(l)) \,|\, \mu_{V_{(\alpha)}}(l) = \mu_V(l) \ge \alpha\}$ – нечеткое множество уровня α с непрерывной функцией принадлежности $\mu_{V_{(\alpha)}}(l)$. Носитель нечеткого множества $V_{(\alpha)}$ в таком случае будет представлять собой интервал действительных чисел $Supp(V_{(\alpha)}) = [c', c'']$, и множество $C = \{c_*, \ldots, c^*\}$ возможных значений

числа классов в искомом нечетком c-разбиении P^* может быть определено как подмножество натуральных чисел, определяемое выражением

$$\left\{\lceil c' \rceil \le c_\ell \le \lfloor c'' \rfloor \mid c_\ell \subset Z^+\right\}, \tag{10}$$

где $\lceil c' \rceil = c_*$ – округление числа c' до ближайшего сверху целого из интервала $[c', c'']$, $\lfloor c'' \rfloor = c^*$ – округление числа c'' до ближайшего снизу целого из интервала $[c', c'']$, а Z^+ – множество положительных целых чисел. Для элементов c_ℓ множества допустимых значений числа нечетких кластеров $C = \{c_*, \ldots, c^*\}$ можно определить значения принадлежности $\mu_{\tilde{C}}(c_\ell) = \mu_{V_{(\alpha)}}(l), l = c_\ell, c_\ell \in C$, так что множество C будет представлять собой носитель нечеткого множества $\tilde{C} = \{c_\ell, \mu_{\tilde{C}}(c_\ell)\}, c_\ell \in C$ с дискретной функцией принадлежности. Значения функции принадлежности $\mu_{\tilde{C}}(c_\ell), c_\ell \in C$ могут интерпретироваться как степени адекватности значений $c_\ell \in C$, так что обобщенный показатель оптимальности числа нечетких классов $\tilde{V}_c(P)$ может быть определен в виде $\tilde{V}_c(P) = V_c(P) \cdot \mu_{\tilde{C}}(c_\ell)$, где $c_\ell \in C$ – число кластеров в P.

Для построения множества одноточечных нечетких множеств $\tilde{C}_e, e = 1, \ldots, k$ можно воспользоваться предложенным в [18] эвристическим D-AFC-TC-алгоритмом нечеткой кластеризации с выбором различных расстояний между нечеткими множествами. Результатом работы D-AFC-TC-алгоритма является распределение $R^*(X)$ объектов исследуемой совокупности X по априори неизвестному числу $c, 2 \le c < n$ нечетких α-кластеров для некоторого вычисленного порога сходства α, позволяющего количественно оценить наименьшую степень сходства объектов в нечетких α-кластерах распределения $R^*(X)$. Полученное в результате работы D-AFC-TC-алгоритма при некотором выбранном расстоянии число c нечетких α-кластеров в $R^*(X)$ может задаваться как $\tilde{c}_e$, а значение порога сходства α полученного распределения $R^*(X)$ – как значение $\mu_{\tilde{C}_e}(\tilde{c}_e)$.

Схема модифицированного FCM-CV-алгоритма

Построение нечеткого множества $\tilde{C} = \{c_\ell, \mu_{\tilde{C}}(c_\ell)\}, c_\ell \in C$ с помощью D-AFC-TC-алгоритма либо его модификаций является этапом, предваряющим построение нечеткого c-разбиения P^*, оптимального в смысле выбранного показателя $V_c(P)$.

Общая схема предлагаемой модификации FCM-CV-алгоритма, основанной на вычислении $\tilde{V}_c(P)$ для всех $c_\ell \in C$, и которую можно обозначить как (m)FCM-CV-алгоритм, выглядит следующим образом.

1. Полагается $c_1 := c_*$ и $c_p := c^*$ **и** значения числа классов c в искомом P^* упорядочиваются следующим образом: $2 \le c_1 < \ldots < c_\ell < \ldots < c_p \le n - 1$.
2. Полагается $\ell := 1$.
3. Вычислять:

3.1. с помощью FCM-алгоритма вычисляется нечеткое c-разбиение $P(c_\ell)$ исследуемой совокупности на c_ℓ классов;

3.2. вычисляется значение показателя $\tilde{V}_c(P_\ell) = V_c(P_\ell) \cdot \mu_{\tilde{c}}(c_\ell)$ для полученного нечеткого c-разбиения $P(c_\ell)$;

3.3. производится проверка условия:
если $\ell < 2$,
то осуществляется переход на шаг 5,
иначе осуществляется переход на шаг 4.

4. Производится проверка условия:
если $\tilde{V}_c(P_\ell) > \tilde{V}_c(P^*)$,
то осуществляется переход на шаг 5,
иначе осуществляется переход на шаг 6.

5. Полагается $P(c_\ell) = P^*$.

6. Производится проверка условия:
если $\ell < p$,
то полагается $\ell := \ell + 1$ и осуществляется переход на шаг 3,
иначе нечеткое c-разбиение P^* является искомым результатом и алгоритм прекращает работу.

Схема (m)FCM-CV-алгоритма построена, исходя из предположения, что в качестве $V_c(P)$ используется коэффициент разбиения (3), так что в случае, когда в качестве $V_c(P)$ используется другой показатель, для которого решение задачи определения оптимального числа классов отыскивается в виде $\min\limits_c(V_c(P)),\ c = 2,\ldots,n-1$, на шаге 4 представленной выше схемы следует производить проверку условия $V_c(P_\ell) < V_c(P^*)$; следует также указать, что (m)FCM-CV-алгоритм требует сохранения в оперативной памяти ПЭВМ только двух нечетких разбиений – вычисленного $P(c_\ell)$ и текущего P^*, вместо множества $\Pi(c_*, c^*) = \{P(c_\ell) \mid \ell = 1,\ldots,p\}$ возможных решений задачи классификации. При $\mu_{\tilde{c}}(c_\ell) = 1, \forall c_\ell \in C$, предложенная схема будет являть собой FCM-CV-алгоритм [7].

Иллюстративный пример

Для проведения вычислительного эксперимента были выбраны изображенные на рис. 1 двумерные данные, предложенные в качестве тестовых К.Г. Луни [19], представляющие собой совокупность 15 объектов $X = \{x_1,\ldots,x_{15}\}$. При проведении вычислительных экспериментов исходные данные были пронормированы в соответствии с выражением

$$x_i^t = \frac{\tilde{x}_i^t}{\max\limits_i \tilde{x}_i^t}, \tag{11}$$

и обработаны D-AFC-TC-алгоритмом с использованием относительного обобщенного расстояния Хемминга, относительного евклидова расстояния и относительной евклидовой нормы [18].

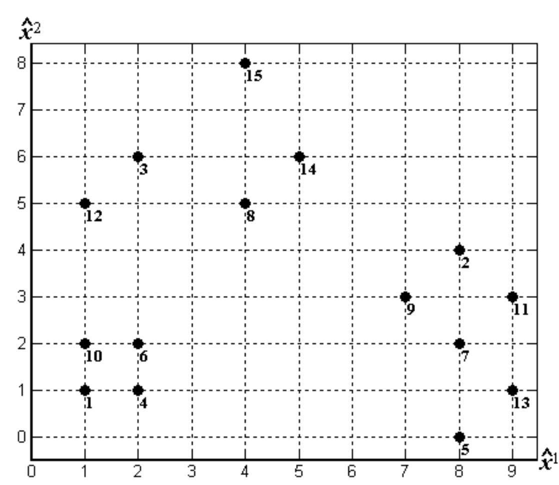


Рисунок 1 – Данные для проведения вычислительного эксперимента

При использовании относительного обобщенного расстояния Хемминга было получено распределение по €$_1 = 2$ при $\alpha_1^* = 0.8125$, а при использовании относительного евклидова расстояния и относительной евклидовой нормы было получено €$_2 =$ €$_3 = 3$ при $\alpha_2^* = 0.8065$ и $\alpha_3^* = 0.9626$ соответственно, что дает возможность построить нечеткие числа $V_1 = (m_1, a_1, b_1)_{LR}$ и $V_2 = (m_2, a_2, b_2)_{LR}$ с модальными значениями $m_1 = 2, \quad m_2 = 3$ и функциями принадлежности

$$\mu_{V_1}(l) = \begin{cases} \dfrac{1}{2} + \dfrac{1}{2}\sin\left(\dfrac{\pi}{2-1}\left(l - \dfrac{1+2}{2}\right)\right), & 1 \le l \le 2 \\ \dfrac{1}{2} - \dfrac{1}{2}\sin\left(\dfrac{\pi}{15-2}\left(l - \dfrac{2+15}{2}\right)\right), & 2 \le l \le 15 \end{cases}, \tag{12}$$

и

$$\mu_{V_2}(l) = \begin{cases} \dfrac{1}{2} + \dfrac{1}{2}\sin\left(\dfrac{\pi}{3-1}\left(l - \dfrac{1+3}{2}\right)\right), & 1 \le l \le 3 \\ \dfrac{1}{2} - \dfrac{1}{2}\sin\left(\dfrac{\pi}{15-3}\left(l - \dfrac{3+15}{2}\right)\right), & 3 \le l \le 15 \end{cases}. \tag{13}$$

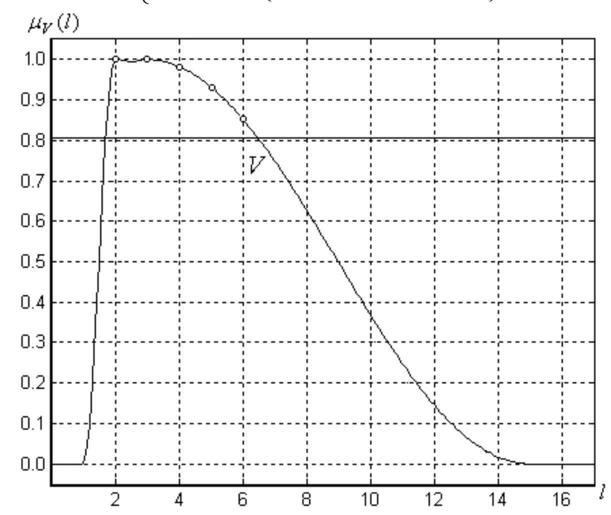


Рисунок 2 – Функция принадлежности нечеткой величины V и значения принадлежностей нечеткого множества V€

Для построения V в качестве S-нормы была выбрана операция объединения нечетких множеств, а значение α было выбрано как $\alpha = \min\limits_{e} \alpha_e^*, e = 1,2,3$ и составило $\alpha_2^* = 0.8065$, так что множество C представляет собой совокупность $C = \{2,3,4,5,6\}$. На рис. 2 непрерывной кривой изображена функция принадлежности $\mu_V(l)$, а символом ○ обозначены значения $\mu_{\tilde{V}}(c_\ell), c_\ell \in C$ нечеткого множества $\tilde{V}$.

На рис. 3 изображено поведение для построенного $\tilde{V}$ обобщенного индекса разделимости $\tilde{V}_{si}(P)$ при обработке данных (m)FCM-CV-алгоритмом.

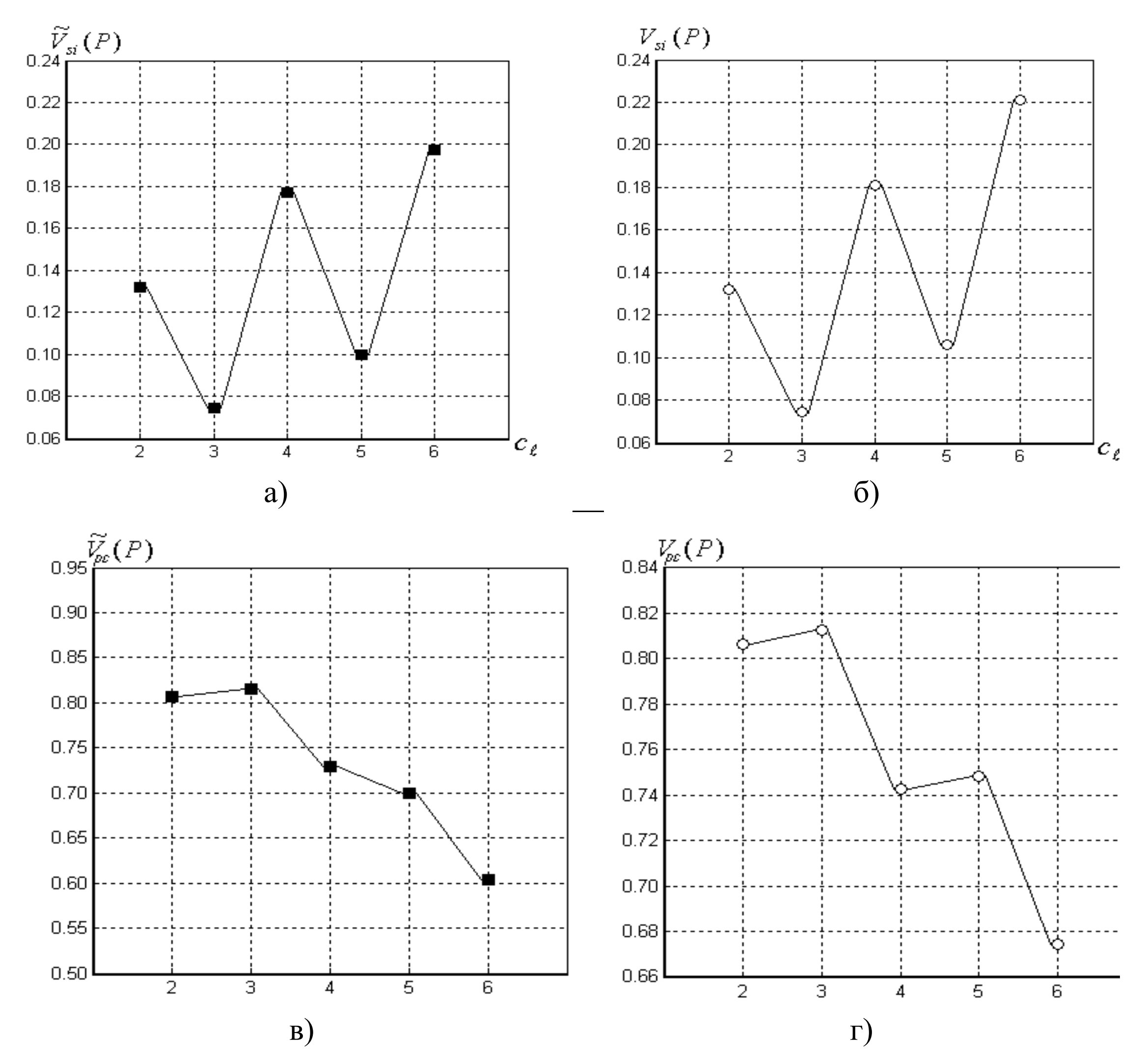


Рисунок 3 – Значения $\tilde{V}_{si}(P)$ (а) и $\tilde{V}_{si}(P)$ при $\mu_{\tilde{V}}(c_\ell) = 1, \forall c_\ell \in C$ (б) при обработке исходных данных (m)FCM-CV-алгоритмом

На рис. 4 изображено поведение для построенного $\tilde{V}$ показателя $\tilde{V}_{pc}(P)$.

В обоих случаях оптимальным числом классов в искомом нечетком c-разбиении является $c = 3$. Значения принадлежностей объектов совокупности $X = \{x_1, \dots, x_{15}\}$ первому нечеткому кластеру оптимального P^* изображены на рис. 4 символом ○, второму – символом Δ и третьему – символом ■.

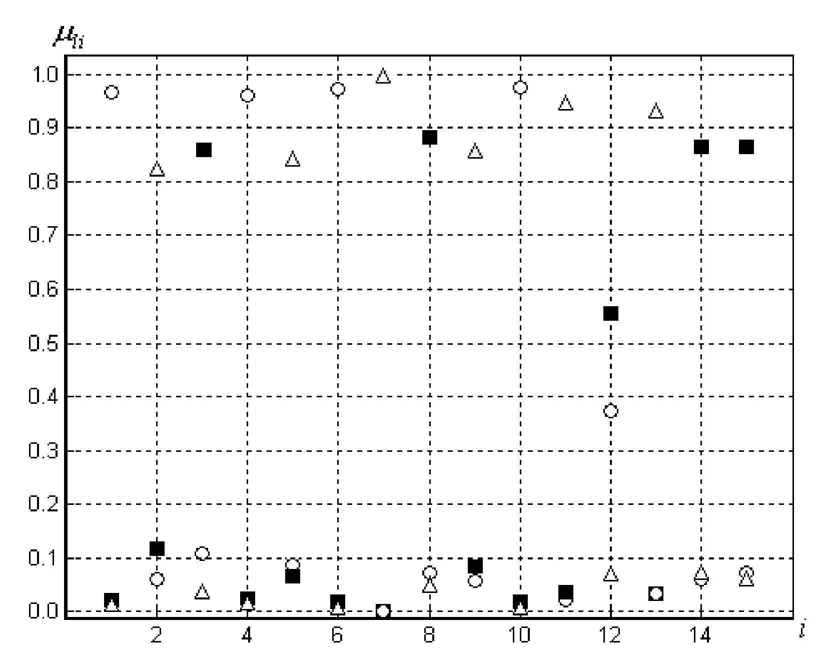


Рисунок 4 – Значения принадлежностей объектов классам оптимального P^*

Анализ исходных данных, приведенных на рис. 1, позволяет визуально выделить три довольно четкие совокупности точек, и приведенные на рисунке 4 значения принадлежности объектов исследуемой совокупности трем классам нечеткого c-разбиения, оптимального в смысле критерия (2), соответствуют относительной близости объектов прототипам нечетких кластеров. Кроме того, результаты проведенного эксперимента показывают корректность определенного в результате работы (m)FCM-CV-алгоритма числа нечетких кластеров в искомом P^*, оптимальном как в смысле индекса разделимости $V_{si}(P)$, так и в смысле коэффициента разбиения $V_{pc}(P)$.

Заключение

Результаты проведенного исследования наглядно демонстрируют, что аппарат нечетких чисел является эффективным средством для построения множества значений наиболее возможного числа классов в искомом нечетком c-разбиении, а метод построения нечетких чисел для решения указанной задачи является достаточно простым и вполне объяснимым с содержательной точки зрения. Кроме того, предложенный подход к построению нечетких чисел на основе задаваемых модальных значений является гибким в том смысле, что функция представления формы нечеткого числа, с одной стороны, а также S-норма для построения нечеткой величины – с другой могут быть выбраны в зависимости от условий задачи.

Так как FCM-алгоритм, отыскивающий минимум функционала (2), представляет собой параметрическое семейство по γ при фиксированном числе кластеров c [2], и при увеличении значения γ возрастает неопределенность классификации, что в свою очередь влияет на поведение показателей $V_c(P)$, то при больших значениях γ иногда оказывается невозможным определить локальный экстремум показателя $V_c(P)$. Этим обстоятельством диктуется целесообразность использования обобщенного показателя $\tilde{V}_c(P)$ в предложенном (m)FCM-CV-алгоритме, анализ данных с помощью которого позволяет в значительной мере снизить требования к оперативной памяти ПЭВМ при

обработке больших массивов данных. В свою очередь, главным достоинством предложенного двухэтапного подхода к решению задачи нечеткой кластеризации, заключающегося в совместном использовании D-AFC-TC-алгоритма и (m)FCM-CV-алгоритма, является возможность обработки данных в полностью автоматическом режиме.

Литература

1. Zadeh L.A. Fuzzy Sets // Information and Control. – 1965. – Vol. 8. – P. 338-353.
2. Прикладная статистика: Классификация и снижение размерности: Справ. изд. / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин; Под ред. С.А. Айвазяна. – М.: Финансы и статистика, 1989. – 607 с.
3. Заде Л.А. Размытые множества и их применение в распознавании образов и кластер-анализе // Классификация и кластер / Под ред. Дж. Вэн Райзина: Пер с англ. / Под ред. Ю.И. Журавлева. – М.: Мир, 1980. – С. 208-247.
4. Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition / F. Höppner, F. Klawonn, R. Kruse, T. Runkler. – Chichester: Wiley Intersciences, 1999. – 289 p.
5. Вятченин Д.А. Нечеткие методы автоматической классификации. – Минск: УП «Технопринт», 2004. – 219 с.
6. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing / J.C. Bezdek, J.M. Keller, R. Krishnapuram, N.R. Pal. – New York: Springer Science, 2005. – 776 p.
7. Вятченин Д.А., Хижняк А.В., Шевяков А.В. Методология построения нечеткого С-разбиения множества объектов на оптимальное число классов // Вестник Военной академии Республики Беларусь. – 2006. – № 4. – С. 16-24.
8. Sun H., Wang S., Jiang Q. FCM-Based Model Selection Algorithms for Determining the Number of Clusters // Pattern Recognition. – 2004. – Vol. 37. – P. 2027-2037.
9. Дюбуа Д., Прад А. Теория возможностей. Приложения к представлению знаний в информатике: Пер. с фр. В.Б. Тарасова. – М.: Радио и связь, 1990. – 288 с.
10. Нечеткие множества в моделях управления и искусственного интеллекта / А.Н. Аверкин, И.З. Батыршин, А.Ф. Блишун и др. / Под ред. Д.А. Поспелова. – М.: Наука, 1986. – 312 с.
11. Вятченин Д.А. Прямые алгоритмы нечеткой кластеризации, основанные на операции транзитивного замыкания и их применение к обнаружению аномальных наблюдений // Искусственный интеллект. – 2007. – № 3. – С. 205-216.
12. Looney C.G. Interactive clustering and merging with a new fuzzy expected value // Pattern Recognition. – 2002. – Vol. 35. – P. 2413-2423.
13. Hathaway R.J., Bezdek J.C., Dawenport J.W. On Relational Data Versions of C-means Algorithms // Pattern Recognition Letters. – 1996. – Vol. 17. – P. 607-612.

Д.А. Вятченін

Використання нечітких чисел задля обґрунтування кластерів у методах нечіткої кластеризації

У статті розглядається метод визначення оптимального числа кластерів у нечіткому c-розбитті, заснований на побудові інтервалу значень найбільш можливого числа класів з використанням нечітких чисел. Запропонована модифікація FCM-CV-алгоритму і наводиться результат обчислювального експерименту.

D.A. Viattchenin

An Application of Fuzzy Numbers for Cluster Validity in Fuzzy Clustering Methods

This paper considers a method of detection of the optimal number of clusters in the fuzzy c-partition based on constructing an interval of values of the most possible numbers of classes with using of fuzzy numbers. A modification of the FCM-CV-algorithm is proposed and a result of the numerical experiment is given.

Статья поступила в редакцию 30.05.2008.